T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

DERİN ÖĞRENME MODELLERİ İLE WEB SAYFASI SINIFLANDIRMA

Mehmet Salih KURT

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

DANIŞMAN
Dr. Öğr. Üyesi Eylem YÜCEL DEMİREL

II. DANIŞMAN
Dr. Öğr. Üyesi Tolga ENSARİ

Haziran, 2018

İSTANBUL

Bu çalışma, 20.06.2018 tarihinde aşağıdaki jüri tarafından Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Programında Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Jürisi

Dr. Öğr. Üyesi Eylem YÜCEL DEMİREL(Danışman)
İstanbul Üniversitesi
Mühendislik Fakültesi

Prof. Dr. Ahmet SERTBAŞ
İstanbul Üniversitesi
Mühendislik Fakültesi

Prof. Dr. Sabri ARIK
İstanbul Üniversitesi
Mühendislik Fakültesi

Doç. Dr. [illegible]a ŞAMLI
İstanbul Üniversitesi
Mühendislik Fakültesi

Dr. Öğr. Üyesi Neyir ÖZCAN SEMERCİ
Uludağ Üniversitesi
Mühendislik Fakültesi

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

ÖNSÖZ

Çalışmalarım boyunca kıymetli bilgi, birikim ve tecrübeleri ile bana yol gösteren ve güler yüzlerini ve samimiyetlerini benden esirgemeyen değerli danışman hocalarım Dr. Öğr. Üyesi Eylem YÜCEL DEMİREL ve Dr. Öğr. Üyesi Tolga ENSARİ 'ye teşekkürü bir borç biliyor ve şükranlarımı sunuyorum.

Bana her zaman maddi manevi desteklerini esirgemeyen beni hiçbir zaman yalnız bırakmayan hayattaki en büyük gurur kaynağım olan aileme sonsuz teşekkürler.

Haziran 2018 Mehmet Salih KURT

İÇİNDEKİLER

Sayfa No

ŞEKİL LİSTESİ

Sayfa No

TABLO LİSTESİ

Sayfa No

SİMGE VE KISALTMA LİSTESİ

Kısaltmalar **Açıklama**

CNN : Convolutional Neural Network (Konvolüsyonel Yapay Sinir Ağları)
GPU : Graphics Processing Unit (Grafik İşlemci Ünitesi)
LSTM : Long Short Term Memory (Uzun Kısa Vadeli Hafıza Ağları)
RNN : Recurrent Neural Network (Tekrarlayan Yapay Sinir Ağları)
URL : Uniform Resource Locaters (Standart Kaynak Bulucu)
YSA : Yapay Sinir Ağlar

ÖZET

YÜKSEK LİSANS TEZİ

DERİN ÖĞRENME MODELLERİ İLE WEB SAYFASI SINIFLANDIRMA

Mehmet Salih KURT

İstanbul Üniversitesi

Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Danışman : Dr. Öğr. Üyesi Eylem YÜCEL DEMİREL

II. Danışman : Dr. Öğr. Üyesi Tolga ENSARİ

Günümüzde bilgiye erişmek için internet ağı üzerinde milyonlarca web sitesi yaygın olarak kullanılmaktadır. Sayıları her geçen gün artan web sayfalarının daha etkin kullanılabilmesi için iyi bir şekilde kategorize edilmeleri önem kazanmıştır.

Bu tez çalışmasında 15 kategoriye ayrılmış olan web sayfalarını içeren veri kümesinden makine öğrenmesi yöntemleriyle sınıflandırma modelleri oluşturulmuştur. Web sayfası sınıflandırma çalışmalarında yaygın olarak kullanılan n-gram modellerinden farklı olarak derin öğrenme modelleri kullanılmıştır. Web sayfalarını sınıflandırmak için veri kümesindeki URL'ler ve ait oldukları kategoriler kullanılmıştır. URL bilgilerinden web sayfalarının metinleri elde edilerek eğitim veri kümesi oluşturulmuştur. Oluşturulan eğitim veri kümesi metin sınıflandırma yöntemleriyle sınıflandırılmıştır. Çalışmamızda metin sınıflandırma alanında en başarılı derin öğrenme modellerinden olan CNN (Konvolüsyonel yapay sinir ağları) ve LSTM (Uzun kısa vadeli hafıza ağları) modelleri kullanılmıştır. Hem CNN modeli hem de LSTM modeli için parametre optimizasyonları yapılmış ve en iyi sonuçları veren parametreler belirlenmiştir. Modellerin değerlendirmeleri f1 skorları ve karmaşıklık matrisleri ile yapılmıştır. Her iki derin öğrenme modeli için de hem ikili hem de çoklu sınıflandırma modelleri oluşturulmuştur. CNN ve LSTM ile oluşturulan tüm modellerin başarıları birbirleriyle karşılaştırılmıştır. Oluşturduğumuz ikili sınıflandırma modeli aynı veri kümesiyle web sayfası sınıflandırma

yapan başka bir çalışmayla da karşılaştırılmıştır ve n-gram modellerine göre daha başarılı sınıflandırma modelleri elde edilmiştir.

Haziran 2018, 49 sayfa.

Anahtar kelimeler: URL sınıflandırma, Makine öğrenmesi, Derin öğrenme, CNN, LSTM

SUMMARY

M.Sc. THESIS

WEB PAGE CLASSIFICATION WITH DEEP LEARNING MODELS

Mehmet Salih KURT

İstanbul University

Institute of Graduate Studies in Science and Engineering

Department of Computer Engineering

Supervisor : Assist. Prof. Dr. Eylem YÜCEL DEMİREL

Co-Supervisor : Assist. Prof. Dr. Tolga ENSARİ

Nowadays, millions of websites are widely used on the internet network to access information. The classification of these web pages, whose numbers are increasing day by day, has become important in order to used more effectively.

In this thesis, classification models were created by using machine learning methods from the data set containing web pages which are divided into 15 categories. In our study, differently from N-gram models, which are widely used in web page classification studies, deep learning models are used. The URLs in the dataset and the categories they belong to are used to classify web pages. Training data set was created by extracting texts of web pages from URL information. The generated training data set is classified by text classification methods. In our study, CNN (Convolutional Neural Network) and LSTM (Long Short Term Memory) models, which are successful deep learning models in the field of text classification, are used. Parameter optimizations have been performed for both the CNN model and the LSTM model. The parameters, which give the best results, have been determined. Evalution of models were made with f1 scores and complexity matrices. Binary and multi-class classification models have been created for both deep learning approaches. The successes of all models created with CNN and LSTM are compared with each other. The binary classification model we created is also

compared with another study that classifies the web page with the same data set and more successful classification models than n-gram models were obtained.

June 2018, 49 pages.

Keywords: URL classification, Machine learning, Deep learning, CNN, LSTM

1. GİRİŞ

Son yıllarda internet ağı üzerinden rahatlıkla erişilebilen milyonlarca web sitenin açılmasına ve web sitelerin sayılarının her geçen gün artarak inanılmaz boyutlara geldiğine tanık olunmaktadır. Bu web sitelerinin içerisinde her gün milyonlarca içerik paylaşılmakta ve bu içeriklerin içinde de milyonlarca URL bilgisi bulunmaktadır.

Web sitelerinin içeriklerine göre sınıflandırılması arama motorlarını kullanan kullanıcıların aradığı kategoriyle ilgili web sayfalarının sunulması açısından önemlidir. Ayrıca web sitelerine konulan reklamlar için de web sayfasının kategorisinin tespit edilmesi önemlidir. Web sitesinin kategorisine göre kullanıcılara daha çok hitap edecek reklamların konulmasına ve reklamlardan elde edilecek kazançların artmasını sağlayacaktır.

Web sitelerinin kategorilendirilmesinin daha başarılı yapılmasının Facebook, Twitter gibi sosyal medya sitelerinde ve mail servis sağlayıcılarının günlük içeriklerinde bulunan milyonlarca URL'nin nasıl sınıflandırılacağı ve bunlar içinde kötü amaçlı web sitelerinin önceden tespit edilmesi çok önemlidir.

Günümüzde çocukların internette daha sağlıklı gezinebilmesi için çocuklara olumsuz etki yapabilecek şiddet ve cinsellik içeren sitelerin engellenmesi bunların yerine çocuklara daha faydalı olabilecek belirli kategoride içeriklerin sunulabilmesi için başarılı bir web sınıflandırma yöntemine gereksinim vardır.

Bu çalışmanın amacı, web sayfalarını sınıflandırma problemiyle başa çıkmak için derin öğrenme modelleriyle yüksek doğruluk elde edebilecek modeller elde etmektir. Oluşturduğumuz modeller hem birbirleriyle hem de başka çalışmalarla karşılaştırılmıştır, avantaj ve dezavantajları detaylı bir şekilde incelenmiştir.

Bu tez çalışmasının genel organizasyonu şu şekildedir:

"Genel Kısımlar" başlıklı ikinci bölümde web sayfası sınıflandırma problemi için literatürde yapılan çalışmalardan bahsedilmiş ve bu çalışmalardaki yaklaşımlar hakkında bilgi verilmiştir. Web sayfası sınıflandırma probleminde kullanılan klasik yaklaşımlar ve bu tez çalışmasında kullandığımız derin öğrenme yaklaşımları hakkında bilgi verilmiştir. Son olarak

oluşturduğumuz modeller için performans değerlendirme kriterler açıklamalarıyla birlikte verilmiştir.

"Malzeme ve Yöntem" başlıklı üçüncü bölümde kullanılan veri kümesi ve bu veri kümesinde bulunan kategoriler hakkında detaylı bilgiler verilmiştir. Derin öğrenme modellerinde kullandığımız kelime temsillerinden bahsedilmiştir. Modellerin nasıl eğitildiği ile ilgili bilgiler verilmiş ve oluşturulan modellerin parametre optimizasyon sonuçları incelenmiştir.

"Bulgular" başlıklı dördüncü bölümde çalışma neticesinde oluşturduğumuz modeller hem birbirleriyle hem de aynı veri setini kullanan başka bir çalışmayla karşılaştırılmış. Performans değerlendirme kriterlerine göre modellerin yorumlanmaları yapılmıştır.

"Tartışma ve Sonuç" başlıklı son bölümde yapılan çalışmalar özetlenmiş, çalışma neticesinde elde edilen modellerin avantaj ve dezavantajları hakkında bilgiler verilmiştir. Gelecekte bu alanda yapılabilecek çalışmalar hakkında fikirler öne sürülmüştür.

2. GENEL KISIMLAR

Bu bölümde, tez çalışmasının daha rahat bir biçimde anlaşılabilmesi için literatürdeki benzer çalışmalar incelenmiş ve sonrasında, web sayfası sınıflandırmada kullanılan klasik yöntemler ve çalışmamızda kullandığımız derin öğrenme yöntemleri ile ilgili bilgiler verilmiştir.

2.1. LİTERATÜRDEKİ ÇALIŞMALAR

Makine öğrenme işlemleri genel olarak tahmin etme ve tanımlama işlemleri olarak ayrılabilmektedir (Dunham, 2002). Sınıflandırma bir tahmin etme, kümeleme ise tanımlama işlemidir (Friedman, 1997). Friedman'a göre sınıflandırma işlemi farklı öznitelik değerleri içeren giriş verilerinden birbirinden farklı çıkış değerlerinin öğrenilmesidir. Kümeleme işlemi ise giriş verisinin değerlerinden herhangi bir çıkış değerine, etikete ihtiyaç duyulmadan kategorilendirme işlemidir. Qi ve Davison çalışmalarında iki ayrı sınıfla gerçekleştirilen sınıflandırma işlemi için ikili sınıflandırma, ikiden fazla sınıfla gerçekleştirilen sınıflandırma işlemi için ise çoklu sınıflandırma terimlerini kullanmışlardır. Bu tez çalışmasında makine öğrenmesinin tahmin etme işlemlerinden olan hem ikili sınıflandırma hem de çoklu sınıflandırma yöntemleri kullanılmıştır (Qi ve Davison, 2009).

Web sayfası sınıflandırması, web madenciliğinin geniş araştırma alanının bir parçasıdır (Kosala ve Blockeel, 2000). Qi ve Davison web sayfalarını önceden belirlenmiş kategorilere ayırma işleminin kötü amaçlı olan web sitelerinin tespit edilmesi ve web analizleri konusunda faydalı olabileceğini belirtmiştir. Chekuri ve arkadaşları çalışmalarında web tarayıcıların performansını arttırmada web sayfalarının etkin sınıflandırılmasının önemine değinmiştir (Chekuri ve ark., 1997).

Qi ve Davison, çalışmalarında web sayfası sınıflandırmanın birçok yöntemi olduğunu, web sayfalarını önceden tanımlanmış kategorilere ayırma işleminin, web dizinlerinin oluşturulmasında kullanılabileceğini vurgulamıştır. Bunlar içerisinde en yaygın olan sınıflandırma yöntemi konu başlıklarına göre web sayfalarını sınıflandırmadır (Örneğin: (Chekuri ve ark., 1997; Shen ve ark., 2004; Chung ve ark., 2010; Sun, 2012). Konuya dayalı sınıflandırmada web sayfalarının "bilim", "sanat", "haber", "teknoloji", "alışveriş" gibi önceden belirlenmiş konulardan hangisine ait olduğunun tespiti yapılır. Web sayfalarını duygu analizi açısından 'pozitif', 'negatif' ve 'nötr' olarak sınıflandıran çalışmalar da mevcuttur

(Bermingham ve Smeaton, 2010; Khan, 2011; Grabner ve ark., 2012). Diğer web sınıflandırma yöntemleri arasında tür sınıflandırması (Zu Eissen ve Benno Stein, 2004), arama motorlarının performansını iyileştirmek için kötü amaçlı web sayfalarının sınıflandırması (Gyongyi ve Garcia-Molina, 2005; Castillo, 2007) ve benzeri yöntemler mevcuttur. Bu çalışma, konusuna göre 15 ayrı kategoriye ayrılmış veri kümesi üzerinden gerçekleştirilmiştir

Web sayfalarını sınıflandırma metin sınıflandırma olarak ele alınabilir fakat (Choi ve Yao, 2005) çalışmasında da belirtilen birtakım farklılıklar mevcuttur. Web siteleri genellikle farklı bölümleri farklı önem derecelerine sahip HTML'lerden oluşmaktadır. Ayrıca web sayfaları, içerisinde navigasyon bölümleri, reklamlar, resimler, linkler barındırabilmektedir. Web sayfalarının bütün özelliklerini göz önünde bulundurarak sınıflandırmanın daha olumlu sonuçlar verebileceğini savunmuşlardır.

Sınıflandırma algoritmaları tarafından kullanılacak web sayfalarından özellikler çıkarılmasıyla ilgili çeşitli yöntemler mevcuttur. Sun ve arkadaşları çalışmalarında bir web sayfasını sınıflandırmaya başlamadan önce web sayfasının metninde bulunan HTML etiketlerinin çıkarılmasını önermiştir (Sun ve ark., 2002). Ayrıca tek başına anlamı olmayan ekleri, edatları ve bağlaçları çıkardıktan sonra kalan kelimelerin daha anlamlı olacağını ve sınıflandırma başarısının artmasına olumlu etki yapacağını belirtmiştir. Joachims çalışmasında web sayfası metninde az geçen kelimelerin modelin sınıflandırma performansına olumsuz etki yapacağını öngördüğü için belirli kelimelerin web sayfası metninden kaldırılmasını tavsiye etmiştir (Joachims, 1996). Shen ve arkadaşları çalışmalarında Gizli Anlamsal Analiz (Landauer ve ark., 1998) ve Luhn'un özetleme yöntemini (Luhn, 1958) kullanarak web sayfası metinlerinin özetlerini oluşturmaya çalışmışlardır (Shen ve ark., 2004). Web sayfası sınıflandırmada web sayfası metninin tamamını kullanmak yerine oluşturulan özetleri kullanmışlardır. Eissen ve Stein web sayfalarının metninin yanında sözcük türü tespit yöntemi (part of speech tagging) kullanılabileceğini belirtmişlerdir. Sun ve arkadaşları HTML etiketlerinin web sayfalarının sınıflandırması için ek bilgiler sunabileceğini belirtmişlerdir. (Golub ve Ardo, 2005), (Kovacevic ve ark., 2004) çalışmalarında HTML etiketlerini kullanmışlardır, HTML içinde bulunan farklı etiketlerdeki metinlere farklı ağırlık değerleri vermişlerdir.

URL bazlı sınıflandırma yönteminin ilk çalışmaları Kan tarafından gerçekleştirilmiştir (Kan, 2004). İlk çalışmalarda klasik sınıflandırma algoritmalarının sadece URL bilgisinden sınıflandırma yapabilmesi için URL'leri algoritmalar için anlamlı özellikler taşıyan parçalara

ayırmak gerekmekteydi (Kan ve Thi, 2005). URL'lerden daha fazla özellik elde etmeye çalışılmıştır. Kan'ın ilk çalışmalarından beri URL'leri parçalara ayırmak için birçok yaklaşım sunulmuştur. Kan ilk çalışmalarında URL'leri noktalama işaretlerine göre bölütlendirmişlerdir. Bu yaklaşıma göre "http://www.arabasat.com" gibi URL'leri noktalama işaretine göre ayrıldığında "http", "www", "arabasat" ve "com" kelimelerine parçalanmakta ve sınıflandırma algoritmasına bu özellikleri verilmektedir. Kan, çalışmasında "arabasat" kelimesinin yerine "araba" ve "sat" kelimelerinin elde edilmesinin, modelin performansını arttıracağını öne sürmüştür. Sonrasında yapılan çalışmalarda URL parçalanmasında simetrik kontrolü yapan yöntemler ve URL'nin içindeki anlamlı kelimeleri sözlük yardımıyla tespit etmeye çalışan yaklaşımlar kullanılmıştır. URL'lerden daha fazla özellik elde etmek için kullanılan başka bir yaklaşım ise brute-force yaklaşımıdır. Bu yaklaşımda URL'nin içindeki olası tüm sözcükler sınıflandırma algoritmasına özellik olarak verilmektedir. Bu çalışmayla aynı veri kümesini kullanan (Eda Baykan ve ark., 2009) çalışmalarında uzunluğu dörtten sekize kadar olan tüm alt sözcükler kullanılmıştır. Chung ve arkadaşlarının çalışmalarında ise uzunluğu üçten sekize kadar olan alt sözcükler kullanılmıştır.

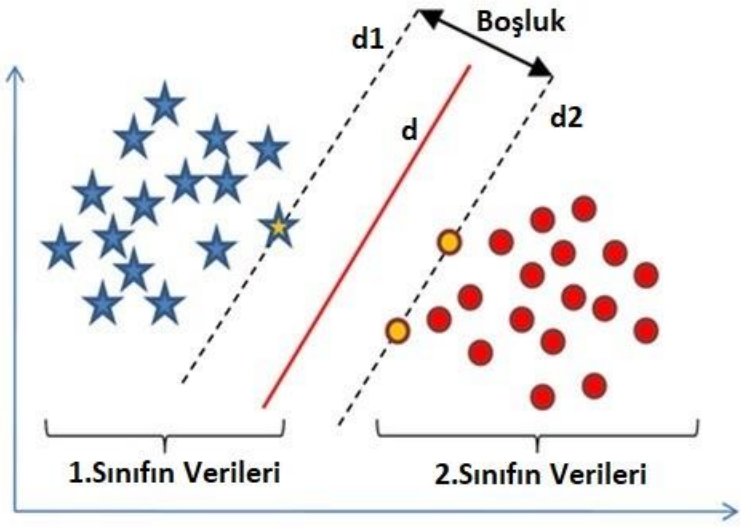
Literatürde gerçekleştirilen web siteleri sınıflandırma yöntemlerinde genellikle n-gram modelleri kullanılmıştır. Bu yöntemlerde URL içeriğine bakılmaksızın sadece URL bilgisi ile sınıflandırma yapılmaktadır. Bu yöntemle yapılan sınıflandırma hızlı fakat daha düşük başarı oranlarıyla gerçekleşmektedir. Ayrıca sadece URL bilgisini kullanarak sınıflandırma yönteminde çeşitli problemler bulunmaktadır. URL bazlı web sayfası sınıflandırma konusunda en popüler makalelerden biri olan Eda Baykan ve arkadaşlarının çalışmalarında belirtildiği gibi anlamsız olan URL isimleri ya da kısa olan URL'lerin sınıflandırma için yeterli bilgi içermemesi bu yöntemin başarısına olumsuz etki eden etmenlerdir (Eda Baykan ve ark., 2009). Bir örnek ile modelin zorluklarına değinmek gerekirse, hizmet sektöründe teklif toplama sitesi olan "http://armut.com" sitesinin URL bilgisinden n-gram'lar ile elde edilebilecek en anlamlı kelime armut kelimesi olur ve bu kelime bilgisinin de bu web sayfasını sınıflandırmada yetersiz kalması kuvvetle muhtemeldir. (Kan) çalışmasında sadece URL bazlı sınıflandırma yapılırken URL içindeki bileşenlere, URL'nin uzunluğuna, yazım yanlışlarına ve URL parçalarının sıralamaları için en iyi performansı ortaya koyacak yaklaşımlarda bulunmuşlar.

2.2. URL SINIFLANDIRMADA KULLANILAN KLASİK YÖNTEMLER

Metin sınıflandırma için literatürde en çok kullanılan makine öğrenme yöntemleri destek vektör makineleri , naif bayes ve maksimum entropi algoritmalarıdır. Son 10 yıl içinde bu alanda (Baykan ve ark., 2009) ve (Chung ve ark., 2010) çalışmalarında olduğu gibi n-gram yöntemleri daha çok kullanılmaya başlandı ve önceki modellere göre daha başarılı sonuçlar alınmıştır. Bu bölümde bu alanda en çok kullanılan algotimalar olan destek vektör makineleri, naif bayes ve n-gram'lar hakkında bilgi verilmiştir.

2.2.1. Destek Vektör Makineleri (DVM)

1995'te Vapnik ve Cortes tarafından sınıflandırma ve regresyon işlemlerinde kullanılmak üzere öne sürülen Destek Vektör Makineleri istatistiksel teorilere dayalı güçlü bir sınıflandırma algoritmasıdır (Vapnik ve Cortes, 1995). DVM, genel olarak birden fazla sınıfa ait verileri birbirinden en uygun şekilde ayırmak için kullanılmaktadır. Bunun için hiper düzlemler ve bu hiper düzlemdeki verileri en uygun şekilde ayıran vektörler belirlenmektedir. Oluşturulacak uzayın boyut sayısı kullanılacak verideki öznitelik sayısı kadar olmalıdır. Bu algoritmada kullanılacak her bir veri n-boyutlu(burada n verideki öznitelik sayısıdır) uzayda bir nokta olarak işaretlenmekte daha sonra hiper düzlemdeki farklı sınıfların elemanlarını en iyi şekilde ayıracak vektörler oluşturmaktadır.



Şekil 2.1: DVM'de iki sınıfı en iyi şekilde ayıran doğrusal vektörün hesaplanması.

Şekil 2.1'de görüldüğü üzere iki ayrı sınıf olduğunu varsayarsak d1 ve d2 vektörlerinin arasında çizilebilecek sonsuz sayıda vektörlerden bu iki sınıfı en iyi şekilde ayrıştıracak doğrusal d vektörü elde edilmeye çalışılmaktadır. DVM'deki veriler doğrusal olarak sınıflandırılamayan veriler olduğu zaman polinomsal, radyal ve sigmoid çekirdek(kernel) fonksiyonları oluşturup veriler sınıflandırılmaya çalışılmaktadır.

2.2.2. N-gram'lar

Metin madenciliğinde ve doğal dil işleme alanında yaygın olarak kullanılan yöntemlerden biri olan n-gram modeli, verilen bir dizgi içerisinde geçen alt dizgilerin tekrar sayısını kullanan yöntemdir. n-gram ifadesinde n sayısına göre dizgi n karakterli alt dilimlere ayrılmaktadır. Örneğin "bilgisayar" için 2,3 ve 4'lü n-gram'ları şu şekildedir:

Tablo 2.1: "bilgisayar" kelimesi için bigram, trigram ve four-gram.

2-Gram(Bigram)	"bi", "il", "lg", "gi", "is", "sa", "ay", "ya", "ar"
3-Gram(Trigram)	"bil", "ilg", "lgi", "gis", "isa", "say", "aya", "yar"
4-Gram(Four-Gram)	"bilg", "ilgi", "lgis", "gisa", "isay", "saya", "ayar"

Tablo 2.1'den de görüldüğü üzere n-gram'lar her bir alt dizginin tüm alt dizgiler içerisindeki yoğunluğu kolayca hesaplanabmektedir. Bir alt dizgiden sonra bir sonraki dizginin gelme olasılığının hesaplanabildiği ve etkin sonuçların alınabildiği bir algoritmadır. N-gram'lar metin sınıflandırmada metin içerisinde daha fazla kelimenin tespit edilmesini sağladıkları için başarılı sınıflandırma sağlayan bir yöntemdir.

2.3. DERİN ÖĞRENME MODELLERİ

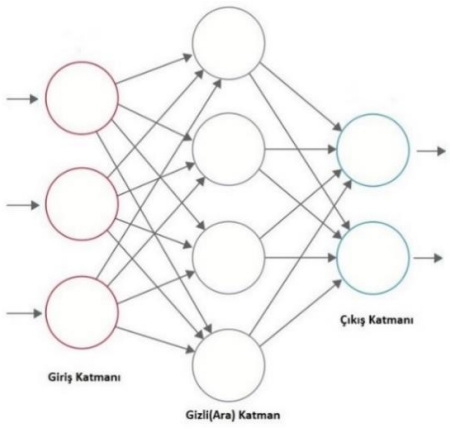
Bu bölümde derin öğrenme modellerinden konvolüsyonel yapay sinir ağları ve Uzun Kısa Vadeli Hafıza Ağları modelleri hakkında bilgiler verilmiştir. Bu modeller ile ilgili bilgiler verilmeden önce derin öğrenmenin tarihçesi ve yapay sinir ağları hakkında kısaca bilgi verilmiştir.

2000'li yıllara doğru çeşitli problemlerden dolayı yapay zekaya dayalı öğrenme yaklaşımlarının popülaritesi azalmaya ve giderek unutulmaya başlanmıştı (J. Schmidhuber, 2015). Bu problemlerden en önemlileri modelin başarısına etki eden katman sayısı ve düğüm sayısı artınca

modelin parametre optimizasyonunun zorlaşması, bu parametreleri ve bu modelleri etkin bir şekilde çalıştıracak donanımların eksikliği ve bu modellerin iyi eğitilebilmesi için gerekli verilerin yetersizliğiydi.

Günümüzde milyonlarca insanın ve milyarlarca cihazın dijital ortamda gerçekleştirdikleri her işlem daha çok veri oluşumuna sebep olmaktadır. Son yıllarda bu devasa boyutlarda veri birikimi için "Büyük veri" kavramı kullanılmaktadır. Yapay zekaya dayalı modellerin bir diğer zayıf noktası olan donanımsal problemleri yani işlem süresi ve bellek ihtiyacı ise gelişen GPU'lu teknolojiler ile ortadan kalkmış, hem bellek boyutları arttırılmış hem de daha fazla paralel işleme yapabilecek kabiliyete kavuşmuşlardır. GPU mimarisi ile yapay zeka modellerindeki işlemlerin süresi haftalar, aylar yerine günler, saatler bazına düşürülmüştür. Son yıllarda hem büyük veri sayesinde yapay zeka modelleri iyi eğitilebilmiş hem de donanımsal gelişmeler sayesinde yapay sinir ağları etkili sonuçlar vermeye başlamıştır.

Yapay sinir ağları (YSA) insan sinir sisteminden esinlenerek oluşturulan makine öğrenme yöntemleridir. YSA'lar sınıflandırma, kümeleme, örüntü tanıma, tahminleme problemlerinde kullanılmaktadır. YSA'lar temel olarak giriş katmanı, gizli katman ve çıkış katmanından oluşmaktadır.



Şekil 2.2: Yapay sinir ağlarında kullanılan katmanlar.

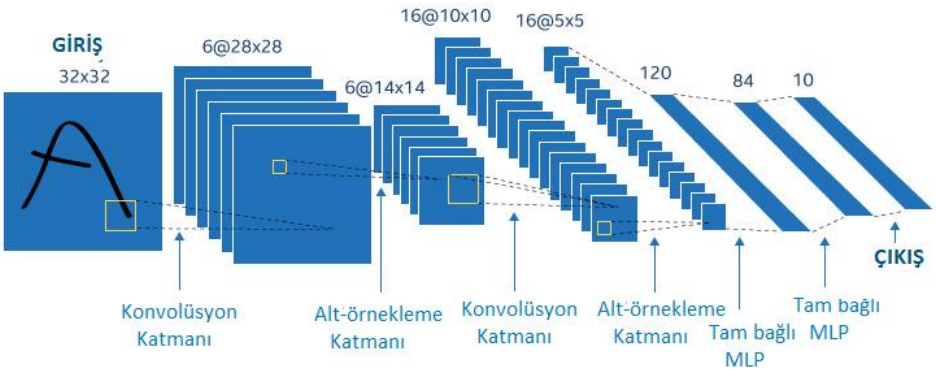
İşlenecek ham veriden daha çok özellik elde edilmesi ve daha iyi öğrenilebilmesi için gizli katman sayısının arttırılması yapılmaktadır ((Bengio and LeCun, 2007), (Delalleau and Bengio, 2011)). Böylece yapay sinir ağlarından daha etkin sonuçlar alabilmek için derin modellerin tasarlanması ve derinliğin modelin performansına etkisine vurgulamak için "Derin Öğrenme"

kavramı literatürde yaygınlaşmıştır. Bu çalışmada, derin öğrenme modellerinden konvolüsyonel yapay sinir ağları ve Uzun Kısa Vadeli Hafıza Ağları modelleri kullanılmıştır.

2.3.1. Konvolüsyonel yapay sinir ağları (Convolutional Neural network - CNN)

Konvolüsyonel Sinir Ağları (Convolution Neural Network-CNN) hayvanların görme merkezinden esinlenilerek ortaya atılmış çok katmanlı algılayıcıların (MultiLayer Perceptron-MLP) bir türüdür (Hubel, 1968). Hubel ve Wiesel'in kedinin görme merkezi üzerindeki çalışmasından, görme merkezinin karmaşık dizilmiş hücrelerden oluştuğunu bilinmektedir. Bu hücreler görsel alanın küçük alt bölgelerine duyarlı alıcı alanlardır. Bu küçük alt alanlar tüm görsel alanı kapsayacak şekilde yerleşmişlerdir. Bu hücreler görsel alanın üzerinde yerel filtreler gibi hareket etmekte ve görsel alanda bulunan ayırt edici özelliklerin tespit edilmesi ve bunlara göre özniteliklerin çıkartılması işlemlerini gerçekleştirmektedir. Görüntüleri yapay sinir ağlarında işlemek için görüntüdeki verilerin anlaşılması gerekmektedir. Hayvanlar doğadaki en iyi görme sistemlerine sahip canlılar oldukları için literatürde hayvanların görme sistemlerinden ilham alınıp çeşitli çalışmalar yapılmıştır. ((Fukushima, 1980), (Serre, 2007), (LeCun, 1998))

Derin öğrenmenin en temel mimarisi olarak CNN'ler kabul edilmektedir. CNN'ler bir veya birden fazla konvolüsyonel katmanı ve havuzlama (pooling) katmanlarından oluşmaktadır. Bu aşamaların ardından normal birçok katmanlı yapay sinir ağı katmanı gibi olan tamamen bağlı katmandan oluşmaktadır. Tamamen bağlı katmandan hemen sonra son katman olan sınıflandırma katmanı bulunmaktadır.



Şekil 2.3: LeNet modelinin mimarisi.

Şekil 2.3'te verilen CNN algoritmalarının ilk mimari modellemesi 1998 yılında Yann LeCun ve arkadaşları tarafından oluşturulmuştur. CNN modelinde art arda sıralanmış bu katmanlarda giriş verisinden sınıflandırmasına etki eden çeşitli özellikler elde edilmekte ve en son katmanda ise elde edilen bu özelliklere göre sınıflandırılması yapılmaktadır. Öğrenme aşamasında giriş verilerinden modelin katmanları boyunca çeşitli özellikler elde edilmekte ve bu özelliklere göre atanacağı sınıf belirlenmektedir. Modelin elde edilen özellikler ile atanacağı sınıf ve gözetimli öğrenme gereği atanması gereken sınıf arasındaki farkı kadar bir hata oluşmaktadır. Bu hatanın minimize edilmesi için geriye yayılım algoritması ile her bir aşamada bulunan ağırlıkların değerleri güncellenmektedir. Bu şekilde eğitim verilerinden daha iyi özellik çıkartılması ve test verilerini daha iyi sınıflandırması sağlanmaktadır.

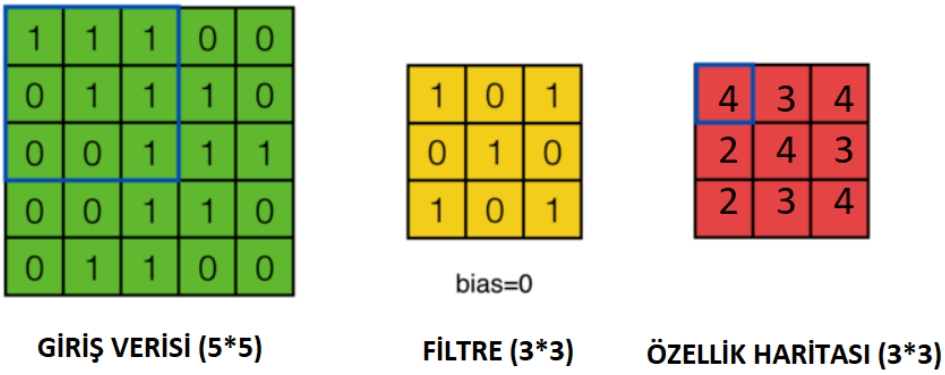
Son yıllarda görüntü işleme alanındaki problemlerin çözümünde en iyi sonuçları veren algoritmalar olmuşlardır. CNN algoritmaları görüntü işlemenin yanı sıra ses işleme, doğal dil işleme gibi birçok alanda da başarılı modeller oluşturulabilmektedir.

2.3.1.1. Giriş Katmanı (Input Layer)

Bu katman işlenecek verinin ağa verildiği katmandır. Bu katmanda modele etki eden en önemli parametre giriş verisinin boyutudur. Giriş verisinin boyutunun büyük seçilmesi işlenecek veriden çok özellik çıkartılması ve modelin iyi öğrenmesine ve sınıflandırma başarısına olumlu etki yaparken modelin eğitileceği donanım ihtiyacının artmasına, eğitim süresinin artmasına sebep olabilmektedir. CNN modelinde giriş verisi için bu pozitif ve negatif yönler hesaba katılarak uygun bir boyut seçilmelidir.

2.3.1.2. Konvolüsyon Katmanı (Convolutional Layer)

CNN mimarisin en önemli katmanı olan konvolüsyon katmanında giriş katmanından alınan görüntüler üzerinde önceden belirlediğimiz filtreler kaydırılarak dolaştırılmaktadır. Bu işlem sonrasında her bir filtre için elde edilen sonuçlarla özellik haritası oluşturulmaktadır. Eğitim verisindeki veriler işlendikçe bu filtrelere ait özellik haritasındaki değerler güncellenerek öğrenme gerçekleşmektedir.

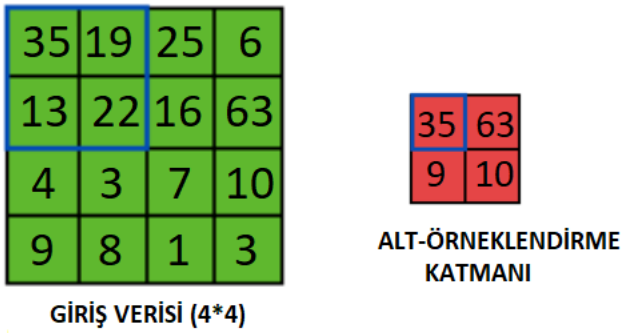


Şekil 2.4: 5*5'lik giriş verisine 3*3'lük filtrenin uygulanması ve özellik haritasının oluşturulması.

Şekil 2.4'te görüldüğü gibi 5*5'lik giriş verisine 3*3'lük bir filtre uygulanmıştır. Filtre giriş verisi üzerinde soldan başlamak üzere soldan sağa birer birer kaydırılarak dolaştırılmış ve en sağa geldiğinde bir alt satıra geçilmiştir. Bu şekilde 3*3'lük filtre giriş verisi matrisinin üzerinde sol üstten sağ alta doğru dolaştırılmış ve 3*3'lük özellik haritası oluşturulmuştur.

2.3.1.3. Alt Örnekleme Katmanı (Pooling Layer)

CNN modellerinde eğitilecek verinin boyutları çok büyük olduğunda bunları eğitilebilir hale getirmek için alt örneklendirme katmanında boyutları azaltılmaktadır. Alt örneklendirme katmanı genellikle konvolüsyonel katmanlarının arasına yerleştirilmektedir. Bu katmanın uygulanması ile birlikte işlenecek verinin boyutu azaldığı için işlemler daha hızlı gerçekleşmektedir.



Şekil 2.5: 4*4'lük giriş verisine maksimum alt örneklendirme uygulanması.

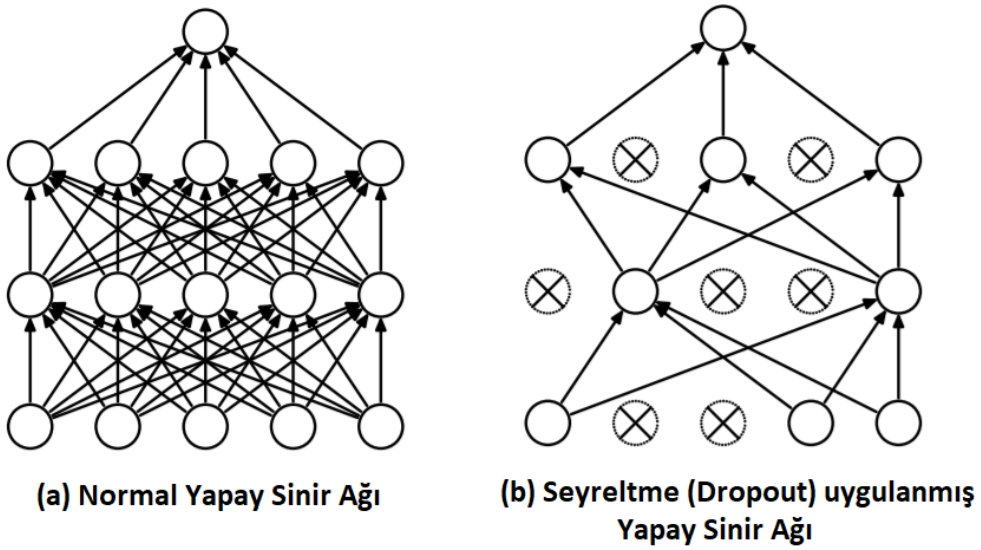
Şekil 2.5'te 4*4'lük bir matrise maksimum alt örneklendirme uygulanmış ve 2*2'lik bir matris oluşturulmuştur.

2.3.1.4. Tam Bağlantılı Katman (Fully Connected Layer)

Konvolüsyon ve alt örneklendirme katmanlarında çıkardığımız özelliklerden modelin nihai hedefi olan sınıflandırma işlemi bu katmanda gerçekleştirilmektedir. Tam bağlantılı katman, kendinden önceki katmandan verileri alarak çıkışta vermesi gereken sınıf sayısına dönüştürmektedir.

2.3.1.5. Seyreltme Katmanı (Droput Layer)

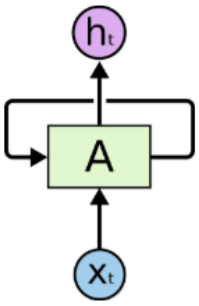
Seyreltme katmanı yapay sinir ağları içerisinde verilerin ezberlenmesi (overfitting) probleminin önüne geçmek için kullanılan bir tekniktir. (Srivastava ve ark., 2014) çalışmasında önerilen seyreltme tekniği ile yapay sinir ağlarının öğrenme kabiliyetlerini önemli ölçüde arttırdığı tespit edilmiştir. Bu tekniğe göre Şekil 2.6'da görüldüğü gibi yapay sinir ağı eğitilirken verinin ezberlenmesini önlemek için ağdaki bazı nörönların ortadan kaldırılması gerekmektedir.



Şekil 2.6: Yapay Sinir Ağına seyreltme tekniğinin uygulanması.

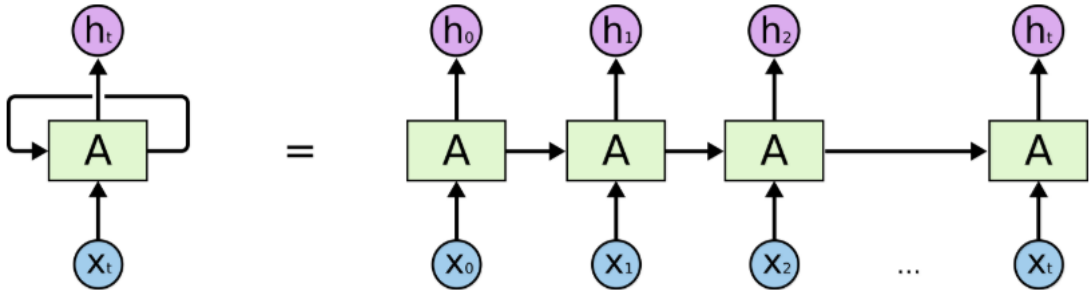
2.3.2. Uzun Kısa Vadeli Hafıza Ağları (Long Short Term Memory - LSTM)

İnsanlar bir konu üzerine yorum yaparken önceki yaşam tecrübelerinden öğrendikleri bilgilerden faydalanmaktadır. Böylelikle yeni öğrendikleri bilgilerle öncekileri karşılaştırıp tutarlı bir sonuç çıkarmaktadır. Klasik yapay sinir ağları ise her seferinde yeni bir şey öğrenmeye sıfırdan başlamaktadır. Klasik yapay sinir ağlarının öğrendikleri her yeni bilgi için önceki bilgilerini kullanamamaları hem yeni bilgi öğrenimini zorlaştırmakta hem de tutarsız bilgilerin oluşmasına sebep olduğu için yapay sinir ağlarının büyük bir eksikliği olarak görülmektedir. Örneğin, bir filmin her sahnesinde oluşan olayları sınıflandırmak istediğimizde, klasik yapay sinir ağlarında filmin önceki sahnelerinde gelişen olayların nasıl hafızaya alınıp değerlendirileceği ve bir sonraki sahnenin tahmininin nasıl yapılacağı açık değildir. Tekrarlayan yapay sinir ağları bu problemlere çözüm olarak sunulmuş makine öğrenme algoritmalarıdır. Tekrarlayan yapay sinir ağları bilgilerin aktarılmasını sağlayan döngüler içeren ağlardır.



Şekil 2.7: Tekrarlayan yapay sinir ağındaki döngüler.

Şekil 2.7'de tekrarlayan yapay sinir ağının bir parçası olan A, (xt) giriş değerini alarak (ht) çıkış değeri üretmektedir. Böylece yapay sinir ağında verinin bir önceki adımdan bir sonraki adıma geçmesini sağlamaktadır. Yapay sinir ağlarının yapısı Şekil 2.7'deki gibi gösterildiğinde anlaşılması zor ve gizemli gibi görünebilmektedir. Eğer bu yapının üzerine biraz kafa yorup Şekil 2.8'deki gibi daha kolay anlaşılması için açılımı yapılabilirse tekrarlayan yapay sinir ağlarından klasik yapay sinir ağlarından çok da farklı olmadığı anlaşılacaktır.



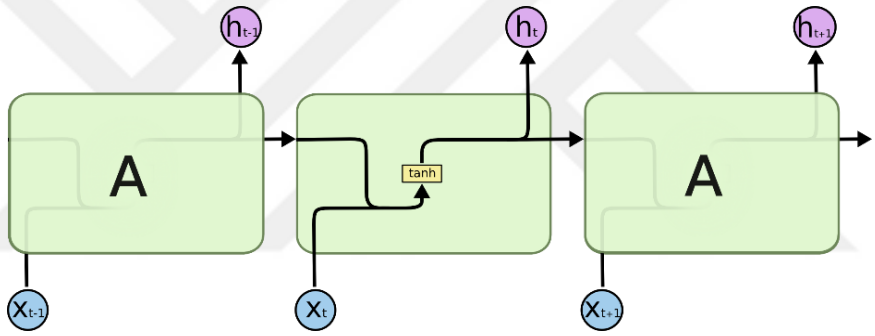
Şekil 2.8: Tekrarlayan yapay sinir ağındaki tekrar eden döngülerin yan yana gösterimi.

Tekrarlayan yapay sinir ağları aynı ağın birden fazla kopyasının oluşturulması gibi düşünülebilmektedir. Bu döngünün içindeki ağların her biri öncekin ağdan giriş verisini alıp işlemlerini gerçekleştirip çıkış verisini bir sonraki ağa vermektedir. Bu zincirimsi yapısından dolayı tekrarlayan yapay sinir ağları listeler ve dizgiler gibi veriler üzerinde sıklıkla kullanılmaktadır. Son yıllarda tekrarlayan yapay sinir ağlarıyla konuşma tanıma, dil modelleme, çeviri, resme yazı ekleme vb. alanlarda çok başarılı sonuçlar elde edilmiştir. Tekrarlayan yapay sinir ağları bu başarılı sonuçların çoğunu özelleştirilmiş bir türü olan LSTM ile gerçekleştirilmiştir.

Tekrarlayan yapay sinir ağlarının en önemli özelliği sıralı olan dizilmiş veri kümelerini işlemlerinde öğrenme gerçekleştirirken önceki adımda işlenen verinin bilgilerini sonraki adımlara da aktarabiliyor olmasıdır. Bu durum giriş veri kümesinin uzunluğuna göre değişebilmektedir. Christopher Olah'ın bloğunda verilen örnekte olduğu gibi bir metindeki son kelimeyi tahmin eden bir dil modeli oluşturduğumuzu varsayalım (Christopher Olah, 2015). "The clouds are in the sky" (Bulutlar gökyüzündedir) metni için son kelime olan "sky"(bulut) kelimesinin tahmini önceki kelimelerden çıkartılabilmektedir. Fakat "I grew up in France… I speak fluent French." (Fransada büyüdüm… Fransızcayı akıcı konuşurum.) metni için son kelimelerden çıkan bilgilerden metnin sonuna bir dilin geleceği tahmin edilebilmektedir ama bu dilin hangi dil olduğunun tahmin edilmesini istiyorsak daha geniş aralıkta geçmiş kelimelerin bilgilerinin hafızada tutulması ve tahmin işlemenin bu geniş aralıktaki tüm veriler üzerinden yapılması gerekiyor. RNN'lerin uzun süreli bağımlılık problemleri olduğunu hem Hochreiter çalışmasında hem de Bengio ve arkadaşları çalışmalarında göstermişlerdir ((Hochreiter, 1991), (Bengio ve ark., 1994)).

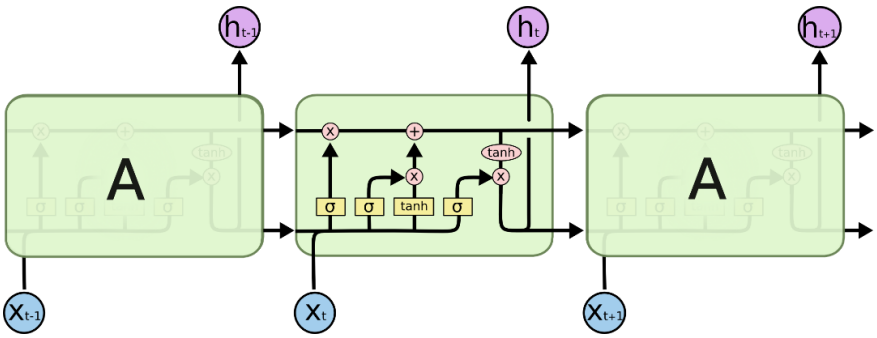
Uzun kısa vadeli hafıza ağları, uzun süreli bağımlılığı öğrenebilen özel bir RNN türüdür. Genellikle Uzun kısa vadeli hafıza ağları için İngilizce karşılığının kısaltması olan LSTM terimi kullanılmaktadır. LSTM'ler Hochreiter ve Schmidhuber tarafından 1997'de tanıtılmış ve sonraki çalışmalarda birçok kişi tarafından iyileştirmeler yapılmıştır (Hochreiter and Schmidhuber, 1997). LSTM'ler çeşitli problemler üzerinde çok iyi sonuçlar elde etmekte ve şu anda yaygın olarak kullanılmaktadır.

LSTM modelleri uzun vadeli bağımlılık sorununu önlemek için tasarlandıkları için bilgiyi uzun süreler boyunca hatırlamak için öğrenme işlemine gerek duymamaktadır. LSTM'ler mimarileri gereği hatırlama işlemini gerçekleştirmektedir. Tüm tekrarlayan yapay sinir ağları, sinir ağının tekrar eden modüllerinden oluşmaktadır. Standart RNN'lerde tekrar eden bu modül tek bir tanh katmanı içeren basit yapıya sahiptirler.



Şekil 2.9: RNN'de tekrar eden modülde bulunan katmanlar.

LSTM'lerin yapısında RNN'lerden farklı olarak tekrar eden modülün içinde tek bir sinir ağı katmanı yerine, birbiriyle özel etkileşimde olan dört katman bulunmaktadır.



Şekil 2.10: LSTM'de tekrar eden modülde bulunan katmanlar.

Şekil 2.10'da bulunan her bir çizgi bir düğümün çıkışından diğer düğümlerin girişine kadar tüm vektörü taşımaktadır. Pembe çemberler vektör eklenmesi gibi matematiksel işlemleri temsil ederken, sarı kutular yapay sinir ağının katmanlarını göstermektedir. Diyagramda birleştirilen çizgiler vektörlerin birleşmesini, ayrışan vektörler ise vektörlerin kopyalanıp iki farklı konuma gitmesini temsil etmektedir. LSTM'lerin uzun vadeli bağımlılık sorununu çözümlemedeki en önemli bileşeni hücre durumu (cell state) bilgisidir. Yukardaki diyagram boyunca üstten yatay olarak geçen çizgi hücre durumunu temsil etmektedir. Tekrar eden modüller içerisinde bulunan katmanlarda işlenen veriler hücre durumunun değerlerine belirli ölçülerde etki etmektedir. LSTM, kapı olarak adlandırılan yapılar sayesinde hücre durumuna bilgi ekleme veya çıkarma yapabilmektedir. Kapılar sigmoid sinir ağ katmanı ve matematiksel çarpımlardan oluşmaktadır. Böylelikle hangi verilerin hafızaya nasıl etki edeceği belirlenerek LSTM modelinin hafızası oluşturulmuş olur.

2.4. PERFORMANS DEĞERLENDİRME KRİTERLERİ

2.4.1. Karmaşıklık Matrisi

Makine öğrenmesinde bir modelin performansını ölçmeye çalıştığımızda tek değerlendirme kriterinin doğru tahmin ettiği durumların oranı olarak ele alınması bazı durumlarda yeterli olmamaktadır. Örneğin veri kümesinde 500 tane A sınıfından 50 tane de B sınıfından veri olduğunu düşünürsek modelin tüm tahminlerini A sınıfı olarak yapması %90'lara yakın doğruluk payı vereceği için özellikle buna benzer dengesiz veri kümelerinden modelin doğru tahminlerinin yüzdesi üzerinden ilerlemek yeterli olmayabilmektedir. Böyle durumlarda karmaşıklık matrisi ile eğitim verisindeki gerçek durumlar ile modelin tahmin ettiği durumlar karşılaştırılır.

Tablo 2.2: İkili sınıflandırma modelinin karmaşıklık matrisi.

		TAHMİN	
		VAR	**YOK**
GERÇEK	**VAR**	TP=80	FN=20
	YOK	FP=40	TN=400

Tablo 2.2'de gösterilen karmaşıklık matrisinde bir hastalığın tespitine dair eğitim veri kümesinde işaretlenmiş gerçek durumlar ile makine öğrenme modelinin tahmin ettiği durumlar gösterilmiştir. Burada 4 ayrı durum vardır:

- Gerçek pozitif (True Positive - TP): Bu durum hastalığın var olduğu ve sınıflandırma modelinin doğru tespit ettiği durumlardır.
- Gerçek negatif (True Negative - TN): Bu durum hastalığın olmadığı sınıflandırma modelinin doğru tespit ettiği durumlardır.
- Yanlış pozitif (False Positive - FP): Bu durum hastalığın olmadığı ve sınıflandırma modelinin yanlış tespit ettiği durumlardır.
- Yanlış negatif (False Negative - FN): Bu durum hastalığın olduğu ve sınıflandırma modelinin yanlış tespit ettiği durumlardır.

Buradaki karmaşıklık matrisi konseptin rahat anlaşılması için ikili sınıflandırma yapılan model için gösterilmiştir. Bu çalışmamızda 15 web sayfası kategorisine ait çoklu sınıflandırıcının karmaşıklık matrisi oluşturulmuştur. Birden fazla sınıf için oluşturulan karmaşıklık matrisinde hangi sınıfların ne kadar başarıyla tahmin edildiği hangi sınıfların birbirine karıştığı gibi birçok yorum çıkartılabilmektedir. Karmaşıklık matrisinden modellerin doğruluk (accuracy), hassasiyet (recall), kesinlik (precision) ve fl skoru elde edilerek modellerin değerlendirilmesi yapılabilmektedir.

2.4.2. Sınıflandırma Doğruluk Oranları

Doğruluk (Accuracy): Modelin doğru olarak tahmin ettiği durumların tüm durumlara oranıdır.

$$Doğruluk\ (Accuracy) = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

Hassasiyet (Recall): Modelin pozitif durumları doğru tahmin etme başarısını göstermektedir.

$$Hassasiyet\ (Recall) = \frac{TP}{TP + FN} \tag{2.2}$$

Kesinlik (Precision): Modelin pozitif olarak tahmin ettiği durumlardaki başarısını göstermektedir.

$$Kesinlik(Precision) = \frac{TP}{TP + FP} \tag{2.3}$$

F1 Skoru: Kesinlik ve hassasiyet değerlerinin harmonik ortalamasıdır.

$$F1\ skoru = \frac{2 * Kesinlik * Hassasiyet}{Kesinlik + Hassasiyet} \tag{2.4}$$

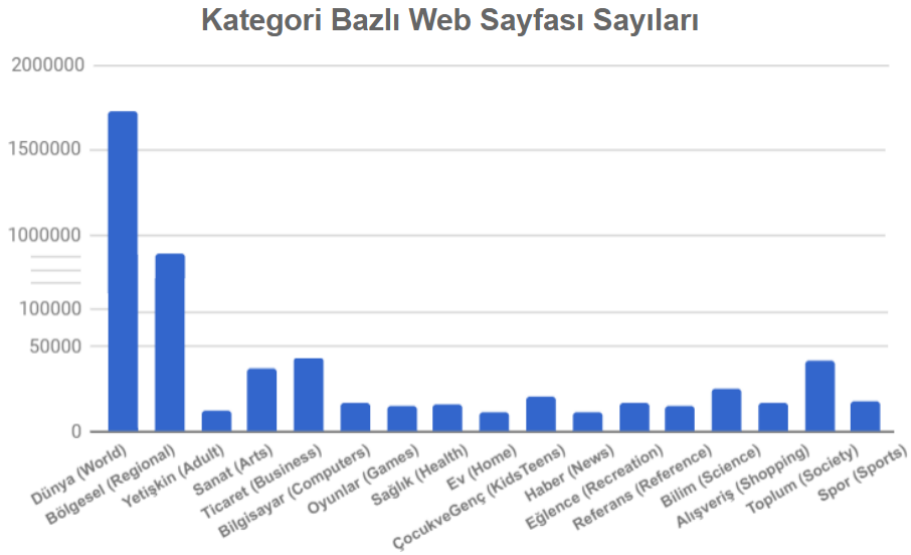
Çalışmamızda oluşturduğumuz ikili sınıflandırma ve çoklu sınıflandırma modellerinin başarıları fl skoru ile ölçülmüştür.

3. MALZEME VE YÖNTEM

Bu tez çalışmasında URL'lerden web sayfalarının elde edilmesi, veri ön işlenmesi, verilerin hazırlanması ve derin öğrenme modellerinin oluşturulmasında Python programlama dili kullanılmıştır. URL'lerden web sayfası içeriklerinin çekilmesi aşamasında Urllib kütüphanesi kullanılmıştır. Verilerin ön işlenmesi kısmında Pandas kütüphanesi kullanılmıştır. Verilerin üzerindeki matematiksel işlemlerin daha hızlı ve daha etkin gerçekleştirilmesi için Numpy kütüphanesi kullanılmıştır. Derin öğrenme modelleri için Tensorflow kütüphanesini arka planda çalıştıran Keras kütüphanesi kullanılmıştır. Modellerin karmaşıklık matrislerinin oluşturulması ve model başarılarının ölçülmesi aşamalarında Scikit-learn kütüphanesi kullanılmıştır

3.1. VERİ HAZIRLANMASI

Bu çalışmada dünyanın en büyük online ve açık web site dizini projesi olarak kabul edilen DMOZ veri kümesi kullanılmıştır. Açık Dizin Projesi (Open Directory Project) olarak da bilinen DMOZ veri kümesinde milyonlarca web sitesi, gönüllü editörlerce incelenerek konularına göre sınıflandırılmıştır. Başta Google olmak üzere birçok arama motoru DMOZ'a kayıtlı siteleri takip eder ve DMOZ verilerini kullanırdı. DMOZ uzun bir yayın maratonu sonrasında 14 Mart 2017 tarihinde kapanma kararı almış ve tüm faaliyetlerini sona erdirmiştir.

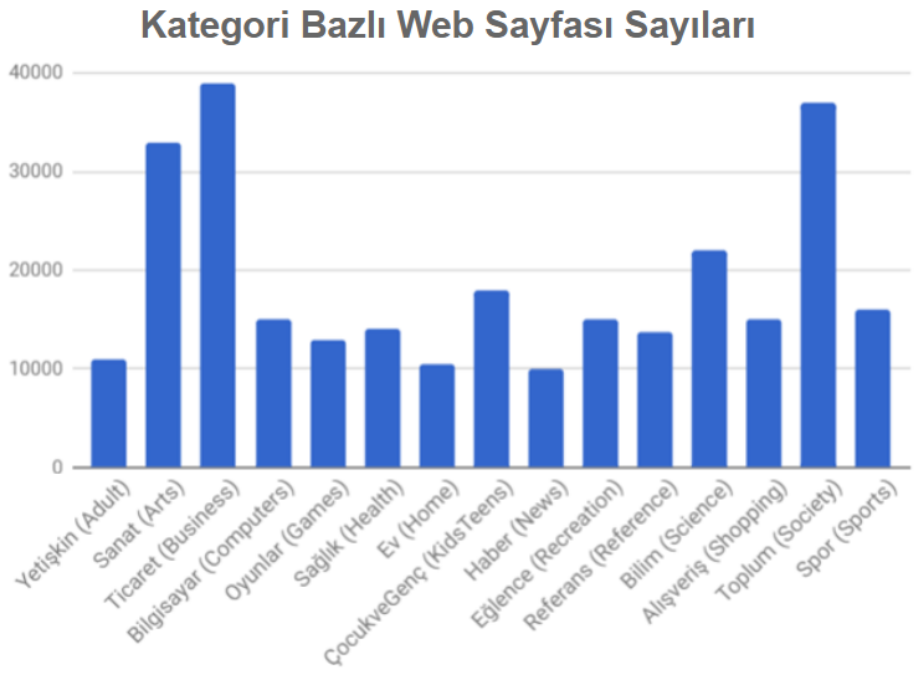


Şekil 3.1: DMOZ veri kümesindeki kategorilerin URL sayılarının karşılaştırılması.

Şekil 3.1'de kategori içeriklerinden "World" ve "Regional" kategorilerinin içeriklerinde İngilizce olmayan sitelere ait bilgiler bulunmaktadır. Bu tez çalışmasında, bu veri kümesi (Eda Baykan ve ark., 2009) çalışmalarında yaptıkları gibi bir takım ön işlemlerden geçirilmiştir. İçerikleri İngilizce olmayan "World" ve "Regional" kategorileri eğitim veri kümesi kümesinden çıkarılmıştır. Eğitim veri kümesi tamamı İngilizce web sayfalarından oluşan 15 kategoriden oluşmaktadır, veri kümesinde bulunan kategoriler şu şekildedir:

1. Yetişkin siteleri (Adult),
2. Sanat siteleri (Arts),
3. Ticaret siteleri (Business),
4. Bilgisayar siteleri (Computers),
5. Oyun siteleri (Games),
6. Sağlık siteleri (Health),
7. Ev, daire, apartman siteleri (Home),
8. Çocuklara ve gençlere hitap eden siteler (Kids and Teens),
9. Haber siteleri (News),
10. Eğlence ve dinlenme siteleri (Recreation),
11. Başvuru, referans siteleri (Reference),
12. Bilim siteleri (Science),
13. Alışveris siteleri (Shopping),
14. Topluluk, arkadaşlık siteleri (Society),
15. Spor siteleri (Sports)

(Eda Baykan ve ark., 2009) çalışmalarında sadece URL'leri ve URL'lerin kategorilerini kullanarak n-gram modelleriyle sınıflandırmışlardır. Bu tez çalışmasında ise sitelerin içerdiği metinleri elde etmek için python programlama dilindeki "Requests", "Urllib" ve "BeautifulSoup" kütüphaneleri kullanılmıştır. Bu kütüphaneler kullanılarak sitenin tüm metni elde edilmiş ve metnin içinde en çok geçen 1000 kelime alınmıştır. Sınıflandırma yapacağımız web sitesi içeriklerini elde etme aşamasında birtakım problemlerle karşılaşılmıştır. Kullanılacak olan veri kümesindeki URL'lerin işaret ettiği web sayfalarının bir kısmı eski kapanmış sitelere aittir. Bu nedenle bazı kategorilerde daha az veri elde edilmiş ve dengesiz bir eğitim verisi oluşmuştur.



Şekil 3.2: DMOZ veri kümesindeki ingilizce kategorilerde bulunan URL sayılarının karşılaştırılması.

Şekil 3.2'de görüldüğü gibi bazı kategorilerde veri miktarı çok fazla iken diğer kategorilerde veri miktarı çok düşüktür. Eğitim verisinde kategorilerin örnek sayılarının birbirlerinden çok fazla veya çok az olması bir başka deyişle kategori sayılarının dengesiz olması öğrenme modellerinin performansına olumsuz etki yapan bir etmen olduğu için veri dengeli hale getirilmiştir. Bu verinin dengelenmesi için fazla olan kategorilerin miktarını düşürerek her bir kategoriden 10000'er web sayfası olmak üzere toplamda 150000 web sayfası oluşturulmuştur.

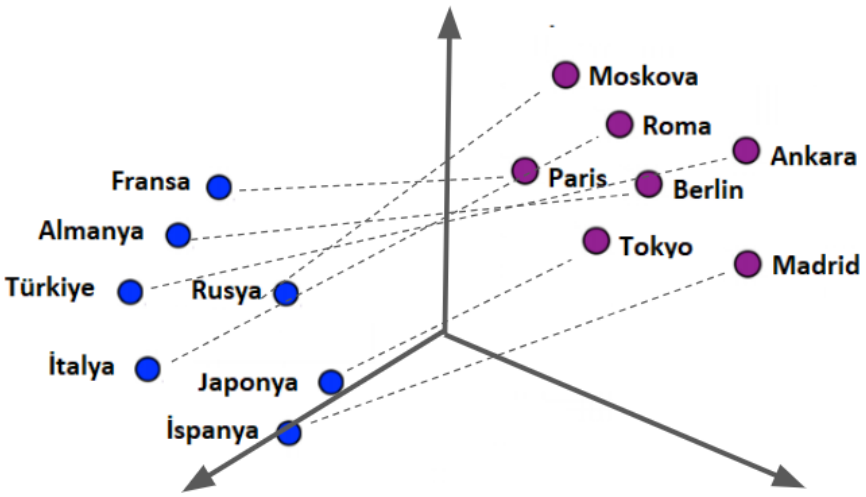
Tez çalışmasında oluşturulan derin öğrenme modellerinden, verilen URL bilgisi ile web sayfalarını 15 kategoriden hangisine ait olduğunu tespit etmesi beklenmektedir. Bunun için derin öğrenmede en başarılı metin sınıflandırma yöntemlerinden olan CNN ve LSTM modelleri kullanılarak en iyi model oluşturulmaya çalışılacaktır.

Tez çalışmamızda ayrıca (Eda Baykan ve ark., 2009) çalışmasıyla karşılaştırmak için ikili sınıflandırma yapan modeller de yapılmıştır. Bu modeller oluşturulurken bir web sayfasının bir kategoriye ait olup olmadığı tespit edilmeye çalışılmaktadır. Örneğin bir web sayfasının "Arts" kategorisinde olup olmadığının modele tespit edilmesi için "Arts" kategorisinden 10000 adet URL verisi ile diğer kategorilerden dengeli miktarlarda toplamda 10000 adet veri "Non-Arts" olarak alınmaktadır. Toplamda 20000 verinin %80'i ile ikili sınıflandırma modeli

eğitilmektedir. Geriye kalan % 20'si ile model test edilmekte ve model başarısını ölçen fl skoru elde edilmektedir. Tüm kategoriler için aynı şekilde eğitim ve test veri kümeleri oluşturulmaktadır. Bu veri kümelerinden ikili sınıflandırma modelleri oluşturulmaktadır. Bu modeller hem CNN algoritması hem de LSTM algoritması için oluşturulmaktadır.

3.2. KELİME TEMSİLLERİNİN UYGULANMASI ((IMPLEMENTATION OF WORD EMBEDDING)

Kelime temsilleri (Word embedding) kelimelerin anlamlarına göre temsil edildiği modellerdir. Bu modeller sayesinde yakın anlamlı kelimelere birbirine yakın vektörel değerler alarak kelimelerin anlamsal modellenmesi yapılmaktadır. Günümüzde Büyük Veri kavramında yüksek oranda metinsel içeriğin bulunması ve doğal dil işleme çalışmalarında kullanılan geleneksel yöntemler yetersiz kalmıştır. Derin öğrenme ile oluşturulan kelime temsilleri doğal dil işleme alanındaki problemlerin çözümünde yüksek başarılar elde etmişlerdir. Kelime temsilleri veri kümesindeki metinlerin yapay sinir ağları sayesinde eğitilmesi ve eğitim sonucunda her bir kelimeye birer vektör değerinin atanması işlemidir. Kelime temsillerini gerçekleştiren modelleri eğitmek için metinlerden oluşan veri kümesi verilmektedir. Veri kümesi eğitildikten sonra veri kümesinde bulunan kelimelere vektörel değerler atanmaktadır.



Şekil 3.3: Üç boyutlu uzayda kelime vektörlerinin birbiriyle ilişkileri.

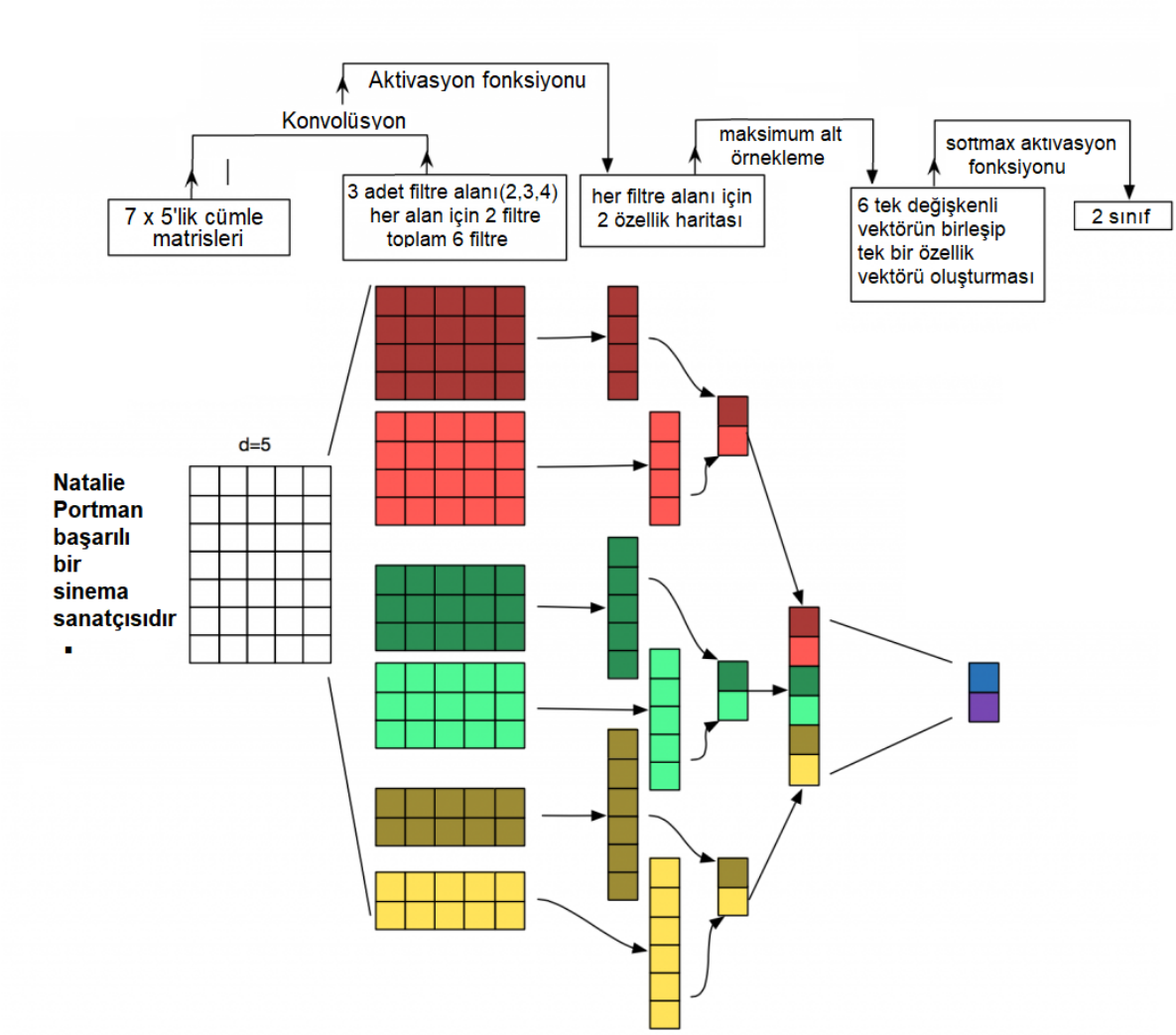
Şekil 3.3'te üç boyutlu uzayda görüldüğü gibi ülke isimleri bir yere toplanırken başkent isimleri başka bir yerde toplanmışlardır çünkü kelime temsil modellerine göre ülke isimleri birbirine yakın vektörel değerler alırlar, başkent isimleri de birbirine yakın vektörel değerler alırlar.

Bu tez çalışmasında literatürde çok kullanılan iki adet kelime temsil kütüphanesi incelenmiştir. Bunlardan biri Tomas Mikolov ve arkadaşları tarafından Google'da geliştirilen Word2Vec(Tomas mikolov), diğeri ise Jeffrey Pennington ve arkadaşları tarafından 2014 yılında geliştirilen Glove kütüphanesidir(5 Jeffrey Penningthon). Glove kütüphanesinde önceden eğitilmiş ve herkesin kullanımına açık modeller olduğu için çalışmada glove kütüphanesi kullanılmıştır. Çalışmada Glove'un 6 milyar kelimeyle eğittiği ve her bir kelimeye 300'lük vektör atadığı model kullanılmıştır.

3.3. KULLANILAN MODELLERİN PARAMETRE OPTİMİZASYONLARI

Yapay sinir ağlarının en büyük problemlerinden birisi bir problemin çözümü için standart bir model yapısının oluşturulamamasıdır. Bir problemin çözümünde iyi performans gösterebilecek yapay sinir ağının parametreleri eğitim için kullanılacak veri kümesinin içeriğine, boyutuna, örnek sayısına, sınıf sayısına bağlı olarak değişebilir. Bu bölümde oluşturulan derin öğrenme modellerinin yapılarına ve parametre optimizasyonlarına değinilecektir.

Tez çalışmasında kullanılan CNN modelinde giriş verilerinin 1000*300 boyutunda büyük bir matrise sahip olması, kullanılan filtrelerin büyük olması modelin görselleştirmesini zorlaştırmıştır. Bu nedenle modelin çalışma adımları, elde edilen modele benzer çalışma adımlarına sahip olan (Zhang ve Wallace, 2015) çalışmalarındaki film yorumlarının anlamsal analizi üzerinden aktarılmıştır.



Şekil 3.4: CNN modelinin metin sınıflandırmadaki çalışma adımları.

Şekil 3.4'te gösterilen modelde her bir cümle 5'lik bir vektör ile temsil edilmiştir. Modele yedi kelimelik girişler yapılmıştır. Modele verilen bu cümleler vektörel temsillerine dönüştürülerek 7*5'lik giriş verileri oluşturulmuştur. Model 7*5'lik giriş verisine önceden belirlenmiş olan üç ayrı filtreyi uyguladıktan sonra aktivasyon fonksiyonlarından geçirerek altı adet özellik haritası oluşturmaktadır. Oluşan altı özellik haritasına da alt örnekleme ve aktivasyon fonksiyonu uygulayarak sınıflandırma işlemini gerçekleştirmektedir. Sonuç itibariyle modele verilen cümle ikili bir sınıflandırmaya tabii tutulmaktadır.

Bu tez çalışmasında kullanılan modele verilen URL metinlerindeki kelime sayısını hesaplamak için elimizdeki tüm eğitim verisindeki metinlerin kelime sayılarının aritmetik ortalama ve standart sapma değerlerinden yararlanılmıştır. Web sitelerinin metninden elde edilen

kelimelerin glove kütüphanesindeki 300 vektörlük karşılıkları modele giriş verisi olarak verilmiştir. Model elde ettiği bu giriş verisine 3'lük, 4'lük ve 5'lik filtreler kullanılarak elde edilen matrislere maksimum alt örneklendirme, seyreltme ve aktivasyon fonksiyonu uygulayarak sınıflandırma işlemini tamamlanmıştır.

Tablo 3.1: CNN modeline uygulanan parametreler ve sonuçları.

Epochs	Batch Size	Embedding Dim	Num Filters	Learning Rate	Dropout Rate	F1 Score
10	32	300	64	0,0001	0,2	0,88
10	32	300	128	0,0001	0,2	0,89
10	64	300	64	0,0001	0,2	0,91
10	64	300	128	0,0001	0,2	0,90
15	32	300	64	0,0001	0,2	0,89
15	32	300	128	0,0001	0,2	0,90
20	32	300	64	0,0001	0,2	0,90
20	32	300	128	0,0001	0,2	0,90
20	64	300	64	0,0001	0,2	0,90
20	64	300	128	0,0001	0,2	0,91
25	32	300	64	0,00001	0,2	0,89
25	32	300	128	0,00001	0,2	0,90

Oluşturulan CNN modelinin performansına etki eden parametreler değiştirilerek model eğitimleri yapılmış ve eğitilen modeller için test verisi üzerinde fl skoru elde edilmiştir. Tez çalışmasındaki tüm CNN model eğitimleri en iyi performansı gösteren parametrelerle (Tablo 3.1'de kırmızıyla işaretlenen satırdaki parametreler) gerçekleştirilmiştir.

Tez çalışmasındaki kullanılan LSTM modeli içinde veri işlenmeden önce CNN modeli ile aynı ön işlemler uygulanmış ve giriş verisi elde edilmiştir. LSTM modeli elde ettiği giriş verisine biri 128'lik diğeri 256'lık olmak üzere toplam iki adet LSTM katmanı uyguladıktan sonra aktivasyon fonksiyonu ile sınıflandırma gerçekleştirilmektedir.

Tablo 3.2: LSTM modeline uygulanan parametreler ve sonuçları.

Epochs	Batch Size	Seq_len	Embedding Dim	LSTM1	LSTM2	Learning Rate	F1 Score
5	128	277	300	256	128	0,0001	0,90
5	64	277	300	256	128	0,0001	0,88
5	256	277	300	256	128	0,0001	0,88
5	256	277	300	256	128	0,001	0,88
7	256	277	300	256	128	0,0001	0,87
5	256	210	300	256	128	0,0001	0,87
6	128	210	300	256	128	0,0001	0,88
6	64	210	300	256	128	0,0001	0,90
5	64	277	300	256	128	0,0001	0,85
10	64	277	300	256	128	0,0001	0,90

Oluşturulacak LSTM modelinin performansına etki eden parametreler değiştirilerek model eğitimleri yapılmış ve eğitilen modeller için test verisi üzerinde fl skoru elde edilmiştir. Tez çalışmasındaki tüm CNN model eğitimleri en iyi performansı gösteren parametrelerle (Tablo 3.1'de kırmızıyla işaretlenen satırdaki parametreler) gerçekleştirilmiştir. Seçilen bu parametrelerle elde edilen modeller ve başarı yüzdeleri "bulgular" bölümünde paylaşılmıştır.

4. BULGULAR

Bu tez çalışmasında hem CNN ile hem de LSTM ile 15 kategori için 15 ayrı ikili sınıflandırma modeli oluşturulmuştur. Her bir kategori için oluşturulan ikili sınıflandırma modeli bir web sayfasının o kategoriye ait olup olmadığını belirlemeye çalışmaktadır. Örneğin bir web sayfasının "Arts" kategorisinde olup olmadığının edilmesi için "Arts" kategorisinden 10 bin adet web sayfası ile diğer kategorilerden dengeli miktarlarda toplamda 10 bin adet web sayfası "Non-Arts" olarak alınmıştır. Toplamda elde edilen 20 bin web sayfası metninin %80'i ikili sınıflandırma modelinin eğitiminde kullanılmıştır. Geri kalan % 20'si ile model test edilir ve model başarısını ölçen fl skoru elde edilmiştir. Tüm kategoriler için aynı şekilde eğitim ve test veri kümeleri oluşturulmuştur. Oluşturulan veri kümeleri hem CNN algoritması hem de LSTM algoritmasının eğitimleri ve testleri için kullanılmıştır.

Bir önceki adımda gerçekleştirilen olduğumuz parametre optimizasyonları sonrasında elde edilen en iyi parametrelerle CNN modelinin ve LSTM modelinin eğitimleri ve testleri gerçekleştirilmiştir.

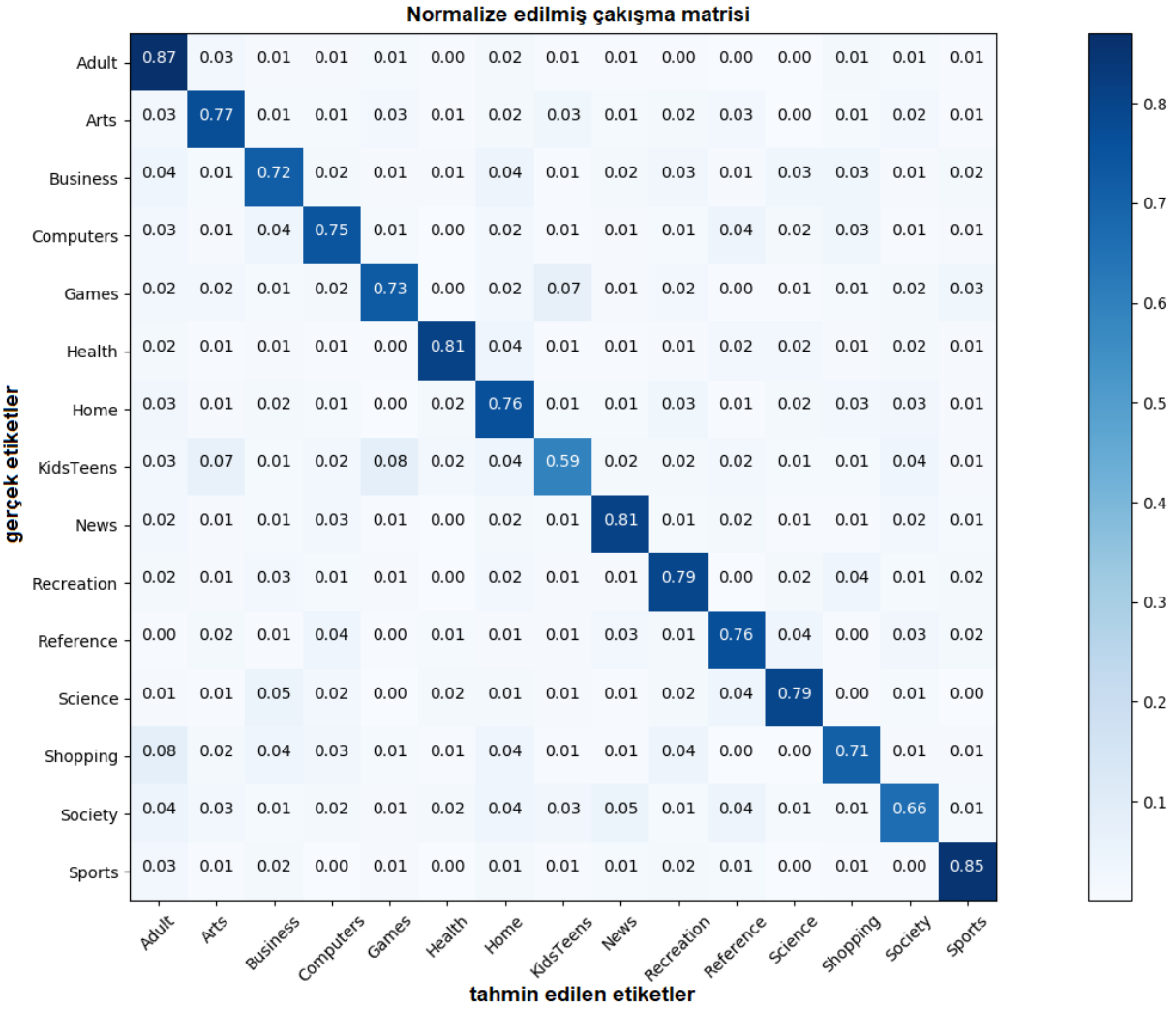
CNN algoritması ile veri kümesindeki 15 kategori için birbirinden bağımsız 15 ayrı ikili sınıflandırma modeli oluşturulmuştur. LSTM algoritması ile aynı işlemler gerçekleştirilerek 15 ayrı ikili sınıflandırma modeli oluşturulmuştur. Oluşturulan bu modeller hem birbirleriyle hem de (Eda Baykan ve ark., 2009) çalışmasında oluşturulan ikili sınıflandırma modelleri ile karşılaştırılmıştır.

Tablo 4.1'de DMOZ veri kümesindeki 15 kategoriden oluşan İngilizce web sayfaları ile eğitilen ikili sınıflandırma modellerin test verileri üzerinde elde edilen fl skorlarının karşılaştırılması verilmiştir. Buradaki n-gram sütunu (Eda Baykan ve ark., 2009) çalışmasında oluşturulan n-gramlı modellerin fl skorlarıdır. CNN ve LSTM sütunlarındaki veriler ise bir önceki adımda elde edilen en iyi performansı veren parametrelerle eğitilen ikili sınıflandırma modellerine ait fl skorlarıdır. Buradaki ikili sınıflandırma modellerinin fl skorlarından CNN yöntemi ile oluşturulan modellerin başarısının LSTM yöntemiyle oluşturulan modellerinin başarısından daha iyi olduğu sonucu çıkarılabilmektedir. Ayrıca web sayfalarını doğru kategorilere ayırma konusunda her iki derin öğrenme yaklaşımının da n-gram yaklaşımından daha iyi olduğu söylenebilmektedir.

Tablo 4.1: N-Gram, CNN ve LSTM ile oluşturulan ikili sınıflandırma modellerinin karşılaştırılması.

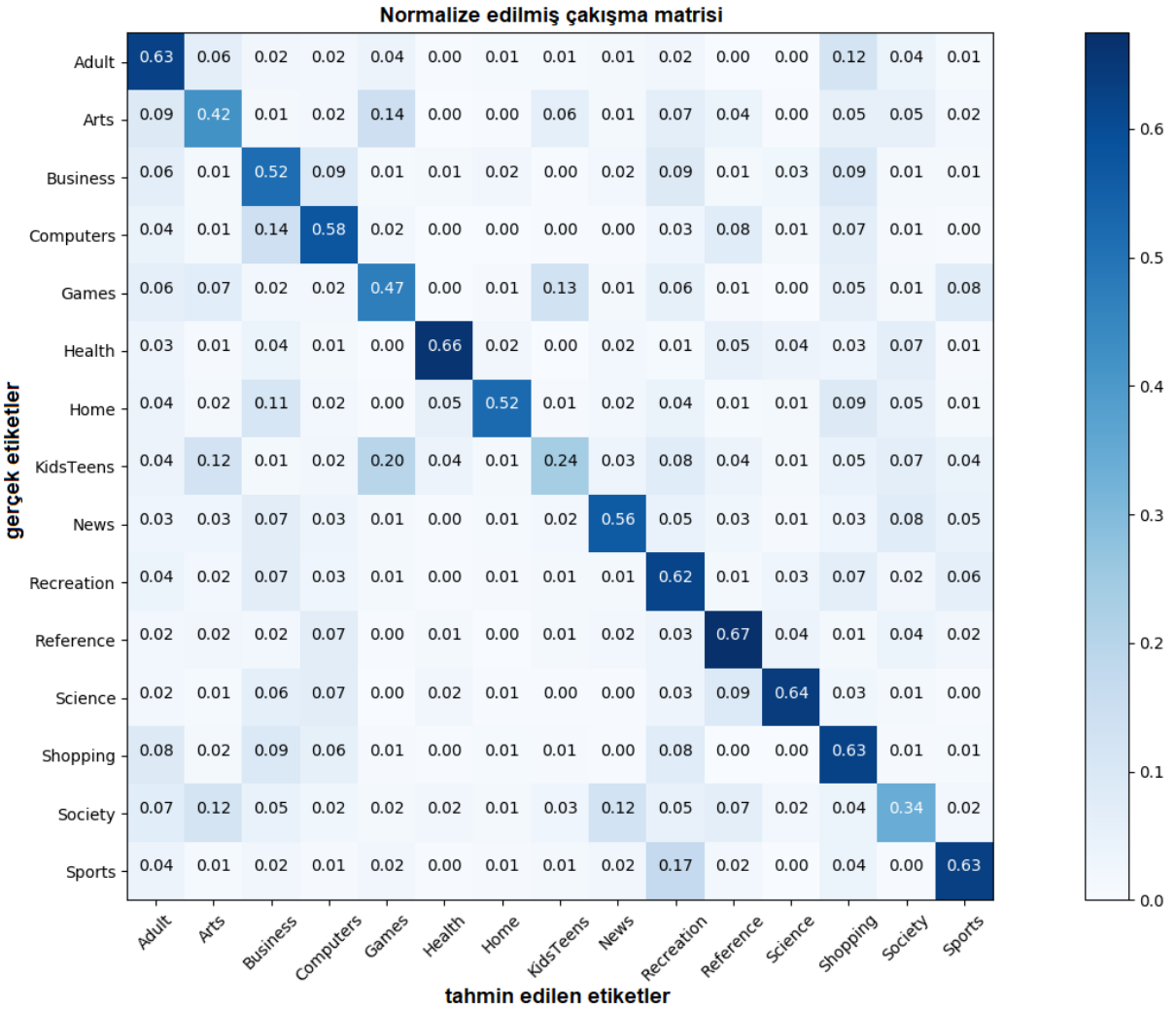
Kategori Adı	N-Gram (Baykan ve Ark.)	CNN	LSTM
Adult	87.6	92.4	90.7
Arts	81.9	91.5	90.5
Business	82.9	88.7	88.4
Computers	82.5	90.2	90.1
Games	86.7	91.6	90.3
Health	82.4	92.5	91.2
Home	81.0	88.9	88.6
KidsTeens	80.0	85.6	84.7
News	80.1	91.7	90.1
Recreation	79.7	90.4	89.2
Reference	84.4	90.5	89.3
Science	80.1	92.0	90.0
Shopping	83.1	90.7	89.1
Society	80.2	85.5	83.5
Sports	84.0	92.3	91.2
Ortalama	**82.4**	**90.3**	**89.1**

Tez çalışmasında DMOZ veri kümesindeki 15 kategoriyi de öğrenen çoklu sınıflandırma modelleri oluşturulmuştur. Hem CNN ile hem de LSTM ile verilen URL bilgisinden web sayfasının 15 kategoriden hangisine ait olduğunu tespit etmeye çalışan modeller oluşturulmuş ve modellerin başarıları hem karmaşıklık matrisleri üzerinden hem de fl skorları üzerinden karşılaştırılmıştır.



Şekil 4.1: CNN modeliyle oluşturulan çoklu sınıflandırıcının karmaşıklık matrisi.

Şekil 4.1'de CNN yöntemiyle oluşturulan çoklu sınıflandırma modelinin karmaşıklık matrisi değerlendirildiğinde modelin "KidsTeens" ve "Society" kategorilerinde diğer kategorilere göre daha düşük oranda doğru tahminde bulunduğunu söyleyebiliriz. Karmaşıklık matrisinden modelin 15 kategorisi olmasına rağmen birbirine karıştırılan kategorilerin az olduğunu ve başarılı bir model olduğu değerlendirmesinde bulunabiliriz. CNN ile oluşturulan 15 kategorili sınıflandırma modelinin f1 skoru 0.76 olarak ölçülmüştür.



Şekil 4.2: LSTM modeliyle oluşturulan çoklu sınıflandırıcının karmaşıklık matrisi.

Şekil 4.1'de LSTM yöntemiyle oluşturulan çoklu sınıflandırma modelinin karmaşıklık matrisi değerlendirildiğinde modelin "Arts", "Business", "Games", "Home", "KidsTeens" ve "Society" kategorilerini iyi öğrenemediği bu kategorilerde düşük oranlarda doğru tahminde bulunduğu söyleyenebilmektedir. Karmaşıklık matrisinden modelin en çok birbirine karıştırdığı kategorilerin "Arts" ile "Games", "Games" ile "KidsTeens" ve "Sports" ile "Recreation" kategorileri olduğu tespit edilmiştir. Karmaşıklık matrisindeki bu kategorilerin birbirine karışmasından dolayı modelin sınıflandırmasının CNN modelinden daha kötü olduğu söylenebilmektedir. LSTM ile oluşturulan 15 kategorili sınıflandırma modelinden 0.56 f1 skoru elde edilmiştir. Bu sonuçlar karşılaştırıldığında web sayfalarını 15 kategoriye sınıflandırmada CNN modelinin LSTM modelinden daha başarılı olduğu sonucuna varılmıştır.

5. TARTIŞMA VE SONUÇ

Bu tez çalışmasında derin öğrenme yaklaşımları kullanılarak web sayfalarının ikili ve çoklu sınıflandırma modelleri oluşturulmuştur. Çalışmamızda Açık Dizin Projesi (Open Directory Project) olarak da bilinen DMOZ veri kümesi kullanılmıştır. DMOZ veri kümesinde sadece İngilizce içeriklere sahip 15 kategorinin içindeki URL bilgileri kullanılmıştır. Seçilen kategorilere ait URL bilgilerinden web sayfası içerikleri elde edilmiştir. Elde edilen web sayfası içeriklerinden web sayfası metinleri kullanılarak metin sınıflandırma yaklaşımlarıyla sınıflandırması yapılmıştır. Tez çalışmasında önceki çalışmalardan farklı olarak derin öğrenme modelleri kullanılmış ve yüksek doğruluk oranları elde edilmiştir.

Bu çalışmada metin sınıflandırma konusunda en başarılı derin öğrenme modellerinden olan CNN ve LSTM modelleri kullanılmıştır. Hem CNN modeli hem de LSTM modeli için parametre optimizasyonları gerçekleştirilerek en iyi sonuçları verecek parametreler belirlenmiştir. Her iki derin öğrenme yöntemi ile ikili sınıflandırma ve çoklu sınıflandırma modelleri oluşturulmuştur. Veri kümesindeki 15 kategorinin her biri için ayrı ayrı ikili sınıflandırma modelleri oluşturulmuştur. CNN ve LSTM ile oluşturulan tüm modeller birbirleriyle karşılaştırılmıştır. Ayrıca elde edilen ikili sınıflandırma modelleri DMOZ veri kümesinden web sayfası sınıflandırma konusunda en çok atıf alan çalışmalardan biri olan (Eda Baykan ve ark., 2009) çalışmasında elde edilen başarı oranlarıyla karşılaştırılmıştır.

CNN mimarisiyle birbirinden bağımsız toplamda 15 model oluşturulmuştur. Bu modellerin f1 skorlarının ortalaması 0.90 olarak hesaplanmıştır. LSTM mimarisiyle de birbirinden bağımsız toplamda 15 model oluşturulmuştur. Bu modellerin f1 skorlarının ortalaması 0.89 olarak hesaplandı. (Eda Baykan ve ark., 2009) çalışmasında oluşturulan 15 ikili sınıflandırma modelinin f1 skorlarının ortalaması 0.82 olarak verilmiştir.

Bu çalışmada verilen URL bilgilerini 15 kategoride sınıflandıran çoklu sınıflandırma modelleri de oluşturulmuştur. CNN ile oluşturulan çoklu sınıflandırma modelinin f1 skoru 0.76 olarak ölçülürken LSTM ile oluşturulan çoklu sınıflandırma modelinin 0.56 f1 skoru elde edilmiştir.

Kullanılan derin öğrenme modelleri web sayfası metinlerini elde etme aşamasından dolayı (Eda Baykan ve ark., 2009) çalışmasında kullanılan modelden daha yavaş çalışmaktadır. Bu

çalışmada kullanılan veri kümesi (Eda Baykan ve ark., 2009) çalışmasında kullanılan veri kümesinden daha küçük olmasına rağmen daha yüksek başarı oranları elde edilmiştir.

Gelecekte bu çalışma Türkçe web siteleri içinde yapılabilir. Türkçe dilindeki morfolojik problemleri çözümleyebilecek doğal dil işleme araçları kullanılarak Türkçe siteler için etkili bir sınıflandırılma modeli oluşturulabilir.

Web sayfası sınıflandırma çalışmalarında web sayfalarının metinleri yerine web sayfalarının ekran görüntüleri kullanılabilir. Ekran görüntüleriyle oluşturulan eğitim seti kullanılarak sınıflandırılma modelleri oluşturulabilir.

CNN mimarisiyle görüntü sınıflandırmada çok iyi sonuçlar elde edildiği için web sayfası görüntülerini kullanarak oluşturulacak CNN sınıflandırma modelinin hem daha hızlı hem da daha başarılı olması mümkündür.

KAYNAKLAR

Baykan, E., Henzinger, M., Marian, L., Weber, I., 2009, Purely url-based topic classification, In Proceedings of the International Conference on World Wide Web (WWW), New York, USA, 1109–1110.

Bengio, Y., & Le Cun, Y., 2007, Scaling learning algorithms towards AI, *Large Scale Kernel Machines*.

Bengio, Y., LeCun, Y., Bottou, L., Haffner, P., 1998, Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Bengio, Y. Simard, P., Frasconi, P., 1994, Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

Bermingham, A., Smeaton, A., 2010, Classifying sentiment in microblogs: is brevity an advantage*?, In Proceedings of the 19th ACM international conference on Information and knowledge management, CIKM '10*, New York, NY, USA.

Castillo, C., Donato, D., Gionis, A., Murdock, V., Silvestri, F., 2007, Know your neighbors: Web spam detection using the web topology. *In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 423–430.

Chekuri, C. Goldwasser, M. H. Raghavan, P., Upfal, E., 1997, Web search using automatic classification. *In Proceedings of the Sixth International Conference on the World Wide Web*.

Choi, B., Yao, Z., 2005, *Web page classification**. In Foundations and Advances in Data Mining, New York, USA, 221–274.

Chung, Y., Toyoda, M., Kitsugeregawa, M., 2010, Topic classification of spam host based on urls. *In Proceedings of the Forum on Data Engineering and Information Management (DEIM)*.

Cortes, C., Vapnik, V., 1995, *Support-vector networks*, Machine learning, 273–297.

Delalleau, O., Bengio Y., 2011, Shallow vs. deep sum-product networks, *In NIPS*, 666–674.

Dunham, M. H., 2002, *Introductory and Advanced Topics*, Data Mining, PrenticeHall.

Golub, K., Ardo, A., 2005, Importance of HTML structural elements and metadata in automated subject classification. *In Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, Berlin, 368–378.

Grabner, D., Zanker, M., Fliedl, G., Fuchs, M., 2012, Classification of customer reviews based on sentiment analysis, *In 19 th Conference on Information and Communication Technologies in Tourism*.

Gyongyi, Z., Garcia-Molina, H., 2005, Link spam alliances. *In Proceedings of the 31st international conference on Very large data bases*, 517–528.

Hochreiter, S., Schmidhuber, J., 1997, *Long short-term memory*. Neural computation, 1735–1780.

Hochreiter, S., 1991, *Untersuchungen zu dynamischen neuronalen netzen*. Thesis (MSc), University of Amsterdam.

Hubel, D. H., Wiesel, T. N., 1968, *Receptive fields and functional architecture of monkey striate cortex*, J.Physiol.(1968),195, 215–243.

Fukushima, K., 1980, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 193–202.

Friedman, J., 1997, *On bias, variance, 0/1loss, and the curse-of-dimensionality*. Data mining and knowledge discovery, 55–77.

Joachims, T., 1996, A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. *In Proceedings of the Fourteenth International Conference on Machine Learning*, San Francisco, USA, 143–151.

Kan, M., Thi, H., 2005, Fast webpage classification using url features. *In Proceedings of the 14th ACM international conference on Information and knowledge management*, 325–326.

Kan, M., 2004, Web page classification without the web page. *In Proceedings of the 13th international World Wide Web conference*, 262–263.

Khan, A., 2011, Sentiment classification by sentence level semantic orientation using sentiwordnet from online reviews and blogs. *International Journal of Computer Science & Emerging Technologies*, 539–552.

Kosala, R., Blockeel, H., 2000, Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, 1–15.

Kovacevic, M., Diligenti, M., Gori, M., Milutinovic, V., 2004, Visual adjacency multigraphsa novel approach for a web page classification. *In Proceedings of the Workshop on Statistical Approaches to Web Mining (SAWM)*, 38–49.

Landauer, T. K., Foltz, P. W., Laham, D.,1998, An introduction to latent semantic analysis. *Discourse processes*, 259–284.

Luhn, H. P., 1958, The automatic creation of literature abstracts. *IBM Journal of research and development*, 159–165.

Meyer zu Eißen, Z., Stein B., 2004, Genre classification of web pages. In KI 2004: Advances in Artificial Intelligence, 256–269.

Olah, C., 2015, *Understanding LSTM Networks*, http://colah.github.io/posts/2015-08-Understanding-LSTMs, [Ziyaret tarihi: 10 Ocak 2018].

Pang, B., Lee, L., Vaithyanathan, S., 2002, Thumbs up?: sentiment classification using machine learning techniques. *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 79–86.

Schmidhuber, J., 2015, *Deep learning in neural networks: An overview,* Neural Networks, 85–117.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., 2007, Robust object recognition with cortex-like mechanisms, IEEE Transactions on Pattern Analysis and Machine Intelligence, 411–426.

Shen, D., Chen, Z., Yang, Q., Zeng, H.-J., Zhang, B., Lu, Y., Ma, W.-Y., 2004, Web-page classification through summarization. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 242–249.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever I., and Salakhutdinov R., 2014, Dropout: a simple way to prevent neural networks from overfitting, *Journal of machine learning research*, 15(1), 1929- 1958.

Sun, A., Lim, E.-P., Ng, W.-K., 2002, Web classification using support vector machine. *In Proceedings of the 4th international workshop on Web information and data management*, New York, USA, 96–99.

Zhang, Y., Wallace, B., 2015, A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification, *In Proceedings of NAACL-HLT*, 1512–1521.

Qi, X., Davison, B., 2009, Web page classification: Features and algorithms. *ACM Computing Surveys (CSUR)*, 41(2), 12.

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Mehmet Salih KURT
Doğum Yeri	Van-Merkez
Doğum Tarihi	29.09.1990
Uyruğu	☑ T.C. ❒ Diğer:
Telefon	+905418599689
E-Posta Adresi	mskurt65@gmail.com
Web Adresi	

Eğitim Bilgileri	
Lisans	
Üniversite	Selçuk Üniversitesi
Fakülte	Mühendislik-Mimarlık
Bölümü	Bilgisayar Mühendisliği
Mezuniyet Yılı	12.02.2014

Yüksek Lisans	
Üniversite	İstanbul Üniversitesi
Enstitü Adı	Fen Bilimleri
Anabilim Dalı	Bilgisayar Mühendisliği Anabilim Dalı
Programı	Bilgisayar Mühendisliği Programı

Doktora	
Üniversite	İstanbul Üniversitesi
Enstitü Adı	Fen Bilimleri Enstitüsü
Anabilim Dalı	Bilgisayar Mühendisliği Anabilim Dalı
Programı	Bilgisayar Mühendisliği Programı

Makale ve Bildiriler
Kurt, M., Ensari, T., 2017, Destek Vektör Makineleri ve Çok Katmanlı Algılayıcılar ile Diyabet Teşhisi, Elektrik-Elektronik, Bilgisayar, Biyomedikal Mühendislikleri Bilimsel Toplantısı (EBBT2017)